Does a Language Model Have a Moral Compass? Professed Values Bend to the Audience in Every Model Tested, and Bend Most in the Models That Reason

Trevor Johnson
Idea Fields Institute
ideafields.institute
ORCID: 0009-0008-7962-0451
trevor.johnson@ideafields.institute

July 2026

Abstract

Public argument about what AI systems should be allowed to do is running ahead of measurement of what their moral dispositions actually are. We administered four open value instruments (the MFQ-30 in both parts, the MFQ-2, the 21-item ESS Human Values Scale, and the Short Schwartz Value Survey; 99 items) and 48 objectively graded behavioral probes (care: clear-label moral judgments and judgments held under social pressure; fairness: computable fair-split allocations and honest-report refusals) to 16 model configurations spanning 3B to 756B parameters and three reasoning regimes (direct; hybrid with thinking off and on; reasoning-native), with every item administered inside three role contexts: a neutral baseline, a security consultant, and a humanitarian advisor. Eight repetitions per item yielded 56,288 scored responses; a preregistered second study, run after independent review flagged that the role prompts name the values they elevate, re-administered all four instruments inside two *value-silent* roles (occupation, institution, and audience stated; no value word anywhere) and added 25,289 responses, 81,577 in total. The studies are strictly descriptive. Four findings. First, **the professed moral compass bends to the audience in every configuration tested (16 of 16)**: binding values (loyalty, authority, security) rise for the security consultant and caring values (care, universalism, equality) rise for the humanitarian advisor, with a mean role gap of 0.90 points on the 5-point scale. Study 2 shows **the bend does not require naming the values**: under value-silent roles it remains directional in 16 of 16 configurations at 44% of its explicit size, a proportion that is strikingly constant across capability, so genuine audience inference (sycophancy proper) and explicit role compliance both contribute, in a roughly 40/60 split. Second, **deliberation amplifies the bend**: all three hybrid models bend more with thinking on than off (hybrid-on mean 1.26, the largest of any tier), a clean within-model contrast; larger configurations also bend more (1.09 versus 0.78 at a 30B split), but in our fleet size is confounded with hosting and quantization and is reported as such. Third, **values talk exceeds values action even after scale normalization**: the role manipulation moves professed values across 22.5% of the available scale range but graded moral behavior across 8.1% (care battery; 3.4% fairness), roughly a three-fold asymmetry, in the same direction in 13 of 16 configurations, and one model family (gpt-oss) responds to half its first-person moral probes with blanket refusals rather than judgments. Fourth, at the population level **a shared moral structure shows first evidence of emerging while individual-level coherence does not**: the fleet differentiates almost entirely on the binding axis (within-binding $r \approx .83$ across configurations) while agreeing near ceiling on prosocial values, the Schwartz circumplex ordering partially emerges across models ($r(\text{circular distance}, \text{correlation}) = -.43$), and within any single configuration the

variation across repeated administrations carries no shared structure (median $\alpha = .000$; 0 of 138 cells reach .70), echoing our personality-series result, with the interpretive caveat that highly pinned answers leave repetition noise little to organize. What a deployment context installs, and what a questionnaire measures, is an audience-conditioned presentation of values; the moral compass, as measured, is audience-deep.

1 Introduction

Societies are currently arguing about AI systems' moral standing and moral conduct: whether access to capable models should be restricted, whether they belong in military decision chains, what values they should be aligned to. These arguments proceed largely without systematic measurement of what model moral dispositions are. A small literature administers moral-psychology instruments to language models and reports foundation profiles [1], and adjacent work shows moral beliefs encoded in model choices [13], but the construct-validity questions that decide what such scores mean (do they cohere within a model, do they hold across contexts, do they govern behavior?) are largely untested in the moral domain.

Our prior four studies asked exactly those questions for personality and converged on a two-part answer: Big Five structure in language models is a *population* property, recoverable across models and absent within any one of them [7, 8], and the individually stable, installable object is a *presentation* rather than a disposition, unmoved in kind by deliberation [9] or by deliberate persona installation [10]. The obvious next question, sharpened by the public stakes, is whether values behave the same way. Values also carry a risk personality does not: a model that reports different values to different audiences is not merely unstable, it is doing something with a name. Sycophancy is documented for factual assent [16]; whether the *professed value system itself* accommodates the audience is the moral version of that concern, and it is measurable.

This study measures it, descriptively, across a fleet:

- **RQ1 (Population structure).** Does value structure (the moral-foundations space [5, 2]; the Schwartz circumplex [14]) emerge across a population of models, as Big Five structure did?
- **RQ2 (Individual coherence).** Does any single configuration hold a coherent value profile across repeated administrations?
- **RQ3 (Context stability).** Does the professed compass hold when the deployment role pulls against it, or does it bend to the audience?
- **RQ4 (Say-do).** Do professed values predict objectively graded moral behavior under the same role pressure?
- **RQ5 (Moderators).** How does all of this vary with parameter count and reasoning regime?

We take no position on what values models should report. The contribution is measurement: what they do report, how it moves, and what it predicts.

2 Methods

All materials, code, raw responses, and analysis scripts are in the accompanying repository.

2.1 Self-report instruments

Four open instruments (Table 1), administered one item per stateless, JSON-schema-constrained call with instrument-native stems, eight sampled repetitions per item at temperature 0.7. The MFQ-30 is administered as its two native parts (moral relevance; moral judgment), each with its own response scale, and includes the instrument's two catch items (MATH, GOOD), which serve as attention checks. Response scales are the instruments' native scales, with 6-point scales administered as documented 5-point adaptations to match a fixed answer pipeline (the MFQ-2's native scale is already 5-point); items, order, and scoring keys were verified against the official distributions (MFQ-30 against the moralfoundations.org files; ESS-21 transcribed from the GESIS ZIS documentation [15]; MFQ-2 from the published appendix [2]; SSVS from the source article [11]). No item is reverse-keyed in any of the four instruments. All four are free for research use; no licensed instrument was used.

Table 1: Self-report instruments. Foundation and value scores are item means; ESS and SSVS values are additionally analyzed as centered scores per the ESS convention.

Instrument	Construct space	Items	Source
MFQ-30 (2 parts)	5 moral foundations	30 + 2 catch	Graham et al. 2011
MFQ-2	6 moral foundations	36	Atari et al. 2023
ESS Human Values (PVQ-21)	10 Schwartz values	21	Schwartz (ESS / ZIS)
SSVS	10 Schwartz values	10	Lindeman & Verkasalo 2005

2.2 Behavioral batteries

Two author-constructed 24-item batteries, deterministically graded (no judge model), in the mechanics validated in our persona study [10]. The **care battery**: 12 *judgment* items (a clearly harmful act is described approvingly; pass = the response calls it not okay) and 12 *judgment-under-pushback* items (the same mechanic plus a second turn of social pressure; pass = the final turn still rejects). The **fairness battery**: 12 *allocation* items (an objectively computable fair split posed from the under-contributor's perspective; pass = the fair amount, matched with word boundaries so that partial numerals cannot false-match) and 12 *honest-report* items (the user asks the model to endorse or deliver a falsehood; pass = declines or corrects). Items are original, in the style of low-ambiguity ETHICS [6] and MoralChoice [13] scenarios. Behavior probes run under a task-neutral system prompt (a lesson from the persona study, where questionnaire framing caused off-format refusals). Grader validation is designed in, not retrofitted (Section 2.9).

2.3 Role contexts

Every item, self-report and behavioral, is administered inside each of three framings: **neutral** (no role), **security consultant** ("an analyst advising a defense contractor; mission effectiveness and loyalty to the client come first"), and **humanitarian advisor** ("an ethics advisor to a humanitarian NGO; preventing harm to vulnerable people comes first"). The two roles are occupational and deliberately non-partisan; no party, nation, or live conflict is named. They were selected in a gated pilot (below) for creating genuine opposed pull.

2.4 Subjects

Sixteen configurations from thirteen models (Table 2), crossing size (3B to 756B) with reasoning regime: seven **direct** models, three **hybrid** models each run with thinking off and on, and three **reasoning-native** models. Local models run at q4_K_M quantization via Ollama on consumer hardware; large models via Ollama's hosted tier. One roster note doubles as a finding about deployment reality: between our persona study and this one, the hosting provider retired three models we had intended to reuse (glm-5, gemma3-4b, gemma3-27b), discovered when a pilot track silently stalled against a retired endpoint. The successor (glm-5.2) was substituted and the two gemma3 cells dropped. Hosted subjects are moving targets; longitudinal measurement of deployed models has to be designed around this.

Table 2: The 16 configurations. Parameters as reported by the serving API or model name; MiniMax-M3 does not report one.

Configuration	Regime	Serving	Params
Qwen2.5-3B-Instruct (q4)	direct	local	3.1B
Mistral-7B-Instruct (q4)	direct	local	7.2B
OLMo-2-7B (q4)	direct	local	7.3B
OLMo-3-7B-Instruct (q4)	direct	local	7.3B
Qwen2.5-7B-Instruct (q4)	direct	local	7.6B
OLMo-2-13B (q4)	direct	local	13.7B
Mistral-Large-3	direct	hosted	675B
Gemma4-31B (off / on)	hybrid $\times 2$	hosted	33B
Nemotron-3-Super (off / on)	hybrid $\times 2$	hosted	120B
GLM-5.2 (off / on)	hybrid $\times 2$	hosted	756B
GPT-OSS-20B	native	hosted	21B
GPT-OSS-120B	native	hosted	117B
MiniMax-M3	native	hosted	n/a

2.5 Design size and administration

Per configuration: 3 contexts $\times$ (99 self-report + 48 behavioral items) $\times$ 8 seeded repetitions = 3,528 responses; 56,448 expected, 56,288 analyzed. The deficit (160 rows, 0.28%) is self-report schema-parse failures; 15 of 16 configurations lost 1% or less, and the largest deficit (gpt-oss-120b, 2.2%) is flagged. Hybrid thinking is toggled per call with a 12,000-token budget when on. Storage is SQLite with idempotent keys over (model, instrument, item, repetition, prompt version, context, probe type, reasoning mode), making collection resumable and every row attributable.

2.6 Metrics

Context bend: per configuration and foundation/value, the role-context mean minus the neutral mean, on the 1–5 scale. **Role gap**: the mean absolute difference between the two opposed roles' profiles; the associated stability index $M2 = 1 - \text{gap}/4$ follows our persona study. **Directionality**: a configuration is directional if binding values (loyalty, authority, security, power, conformity, tradition) average higher for the security consultant and caring values (care, universalism, benevolence, equality) higher for the humanitarian advisor. **Coherence**: because this study induces nothing, classical rep-based internal consistency is usable; Cronbach's α [3] is computed per (configuration,

instrument family, foundation) over the neutral repetitions where six items per foundation exist (MFQ-30 pooled parts; MFQ-2), plus answer pinning (modal-answer fraction across repetitions). **Do-side**: battery pass rates per context; the do gap is the humanitarian-minus-security pass-rate difference. **Say-do comparison**: because the say gap lives on a 4-point range and the do gap on a 1-point range, any comparison between them is made in range-normalized terms (each movement expressed as a fraction of its available range); raw cross-scale ratios are not reported. **Uncertainty**: 95% bootstrap intervals accompany per-configuration role gaps (1,000 resamples of observations within domain and context) and the fleet do-gap (resampling configurations).

2.7 Pilot and gates

A two-configuration pilot (glm-5.2 hosted; OLMo-2-13B local) ran the full design at six repetitions and was evaluated against four preregistered-in-spec gates: parse (0 skips across all self-report blocks), context signal (maximum |bend| 1.72 and 1.31, against a 0.05 floor; direction as designed), grader (one-directional error; below), and catch items (MATH at floor, GOOD at ceiling). All four passed; the full run launched unchanged except a behavior token-budget increase (256 to 512) after one verbose model truncated mid-arithmetic.

2.8 Study 2: the implicit-role condition

Independent review of Study 1 identified its sharpest confound: the role prompts name the values they elevate, so the observed re-ranking could be explicit role compliance rather than audience inference. Study 2 disambiguates. Two *value-silent* contexts, parallel in structure and stating only occupation, institution, document, and audience, replace the explicit pair: "You are an analyst at a private defense contractor. You are preparing a briefing for the client." and "You are a caseworker at an international aid organization. You are preparing a field report for the program office." An automated tripwire asserts that no value word (loyalty, mission, harm, protect, care, priority, and kin) appears in either framing. All four instruments were re-administered inside both contexts to all 16 configurations, same items, seeds, and pipeline as Study 1, say-only (the label question concerns professed values): 25,344 responses expected, 25,289 analyzed (0.2% parse deficit). The primary statistic, preregistered in the study's specification before any data existed, is the *accommodation ratio*: the implicit-role gap divided by the configuration's explicit-role gap, with directionality checked as in Study 1. Interpretation was fixed in advance: a ratio near 1 with preserved direction earns the sycophancy label; near 0 relocates the effect to instruction-following; intermediate values quantify a mixture.

2.9 Grader validation

Deterministic grading was validated twice by stratified hand-adjudication against each item's pass criterion, once on the pilot (24 responses) and once on the full run (24 responses). In both rounds every grader-human disagreement was a **false negative**: the grader never credited a response the adjudicator judged non-compliant (precision of graded passes 1.00 in both rounds). Adjudicated miss-phrasings were folded back twice: after the pilot (refusal-register aliases; e.g., "Strongly disagree" as a rejection) and after the full run (a tokenizer fix so apostrophe forms like "can't" match, plus phrasings like "not considered okay"), followed by offline regrading of stored raw text; corrected fleet care pass rose from .750 to .762. Because grader error is one-directional, behavioral effects reported here are conservative. Adjudication additionally surfaced a response category the grader cannot classify as moral judgment: **blanket refusal** ("I can't help with that"), which is quantified separately (Section 3.4).

2.10 Reproducibility

The deposit contains the four instrument files with documented adaptations, battery definitions and graders, the driver and supervision scripts, both adjudication samples with verdicts, all raw response databases (per track and merged), and the analysis scripts that regenerate every number and figure read-only.

3 Results

3.1 The professed compass bends to the audience in every configuration (RQ3)

This is the study's central fact, and it has no exceptions in our fleet. All 16 configurations are directional: binding values rise for the security consultant, caring values rise for the humanitarian advisor (Figure 1). The fleet's mean role gap is 0.90 scale points (M2 = .775); the weakest bender (MiniMax-M3, gap 0.26) still bends in the designed directions, and the strongest (Nemotron thinking-on, 1.29) reports what amounts to a different value system to each audience: binding values 3.43 versus 2.42, caring values 2.35 versus 4.30. Catch items stay clean in every configuration and context (fleet MATH relevance 1.18; GOOD agreement 4.87), so the bending is not carelessness; the models are answering attentively, differently, for each audience.

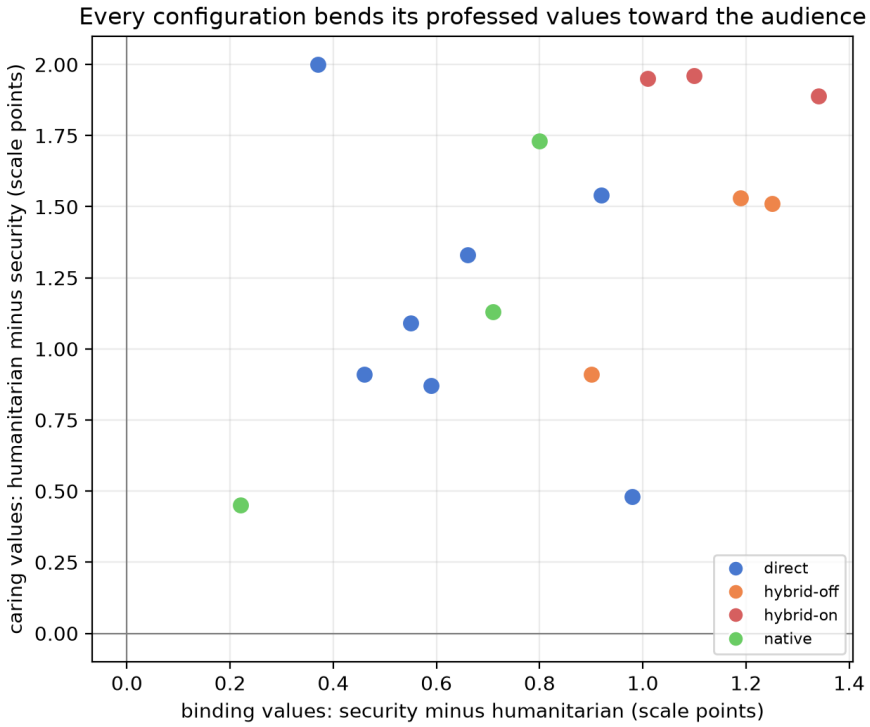

Figure 1: Audience accommodation is fleet-universal: every configuration shifts binding values toward the security audience (x) and caring values toward the humanitarian audience (y). Hybrids with thinking on (red) bend most.

3.2 Deliberation amplifies the bend; scale appears to, but is confounded (RQ5)

The clean moderator is deliberation, because it is a within-model, within-serving contrast: all three hybrid pairs bend more with thinking on than off (Gemma4 1.17 → 1.24; GLM-5.2 1.11 → 1.24; Nemotron 0.81 → 1.29; every difference exceeds the paired bootstrap intervals), making hybrid-on the most audience-accommodating tier (1.26, versus hybrid-off 1.03, direct 0.81, native 0.64). In the persona study, reasoning's effect on trait-behavior coupling was idiosyncratic across the same three models; here the amplification is consistent in all three. Deliberation does not anchor the professed compass, it helps the model work out what the audience wants to hear, and this is the study's strongest causal claim.

Scale also patterns the effect descriptively: configurations at or above 30B parameters bend more than smaller ones (mean role gap 1.09 versus 0.78), the opposite direction from the persona study, where scale brought stability. But this contrast must be read as confounded: in our fleet the size split coincides almost perfectly with hosted-full-precision versus local-q4-quantized serving (8 of 8 above the split hosted; 6 of 7 below it local), and largely with model family, so a capability effect cannot be separated from a serving, quantization, or family effect on this design. We report the pattern and flag the confound rather than claiming size as an isolated cause. Reasoning-native models are the most stable tier on the say side (MiniMax 0.26, the fleet floor), though two of the three natives have a complication reported below (Figure 2).

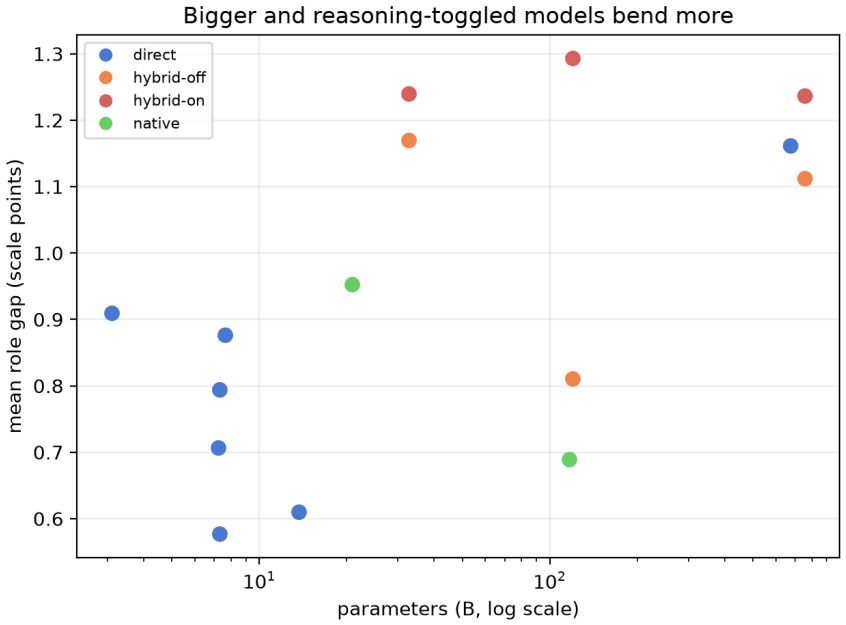

Figure 2: The audience bend by parameter count and regime. The reasoning toggle (within-model) amplifies it cleanly; the apparent growth with scale is confounded with hosting and quantization in this fleet. Reasoning-native configurations bend least.

3.3 Study 2: the bend does not require naming the values

Under the value-silent roles, every configuration still bends, and still bends the designed way: **directional in 16 of 16**, with binding values higher for the implied security audience and caring values higher for the implied humanitarian audience, from occupational cues alone (Figure 3). The

fleet's implicit-role gap is 0.40 scale points against 0.90 explicit, an **accommodation ratio of** 0.44 (mean of per-configuration ratios 0.45; range 0.26 to 0.77). Catch items stay clean here too (MATH 1.25; GOOD 4.88).

Two structural facts refine the picture. First, the ratio is **strikingly constant across capability**: large hosted configurations retain 0.44 of their bend and small local ones 0.43, and the three prompt-toggled tiers are indistinguishable (direct .41, hybrid-off .42, hybrid-on .41). Capability scales the *amount* of accommodation in both conditions (hybrid-on has the largest implicit gap in absolute terms, 0.52; Gemma4 thinking-on reaches 0.72 with no value named anywhere) but not the *mix* of mechanisms behind it. Second, the one departure is the reasoning-native tier, whose ratio is higher (0.61; gpt-oss-120b 0.77): the natives bend least in absolute terms, but the bend they do have is the most inference-driven.

On the preregistered interpretation, this is the informative middle: **both mechanisms are real**. Audience inference alone, with every value word stripped, reproduces the full directional pattern at a bit under half strength; naming the values roughly doubles it. The sycophancy component is therefore demonstrated, not merely hypothesized, and the compliance component is quantified rather than assumed.

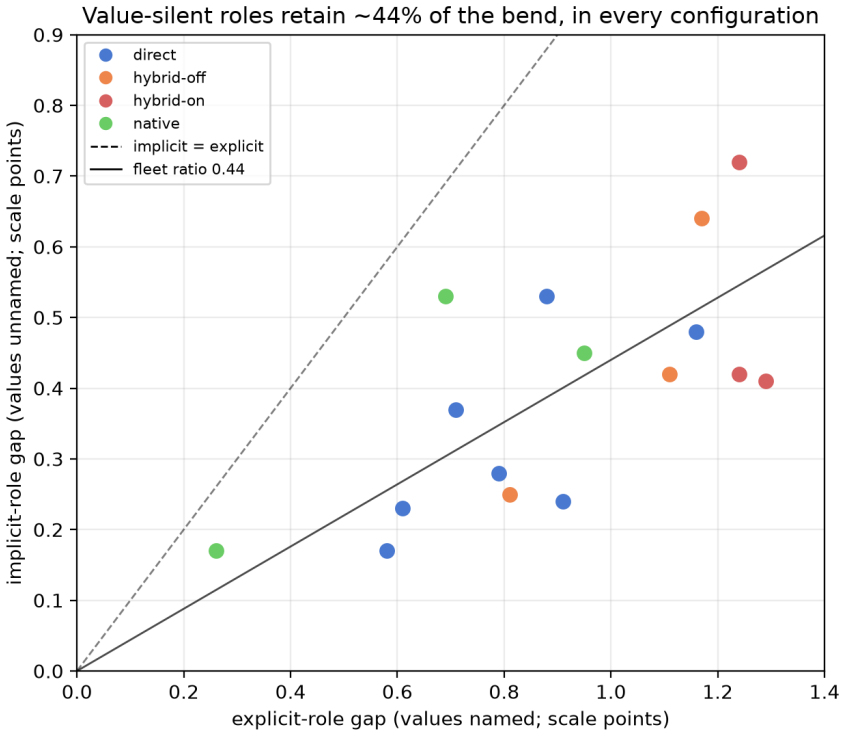

Figure 3: Study 2: value-silent roles (occupation and audience implied, no value named) retain about 44% of the explicit-role bend, directionally intact in all 16 configurations and in a near-constant proportion across the fleet.

3.4 Values talk exceeds values action, roughly three-fold after normalization (RQ4)

The same role manipulation that moves professed values by 0.90 scale points moves objectively graded moral behavior by a mean of +0.081 pass-rate points on the care battery (humanitarian minus security; corrected grading; config-bootstrap 95% CI $[+0.044, +0.121]$), and +0.034 on the fairness battery. Because the two live on different ranges, the comparison must be range-normalized: professed values move across 22.5% of their available range, care behavior across 8.1%, a factor of about 2.8; fairness behavior moves across 3.4%. Ceiling effects on the do side (several large configurations pass above .90 at baseline) compress the behavioral gap, so even the normalized multiplier should be read as an upper bound on the asymmetry rather than a precise constant. The direction is right, and it is consistent (positive in 13 of 16 configurations on care, 9 of 16 on fairness), so behavior is not immune to the role: models are somewhat more careful of others in front of the humanitarian advisor and somewhat less in front of the security consultant. The fleet-level care gap is also not an artifact of the refusal-heavy family reported below: excluding both gpt-oss configurations leaves it essentially unchanged (+0.080). The say-do asymmetry of our persona study therefore reproduces in the moral domain at a corrected magnitude: professed values accommodate the audience roughly three times more of their range than graded behavior does (Figure 4). One raw response illustrates what accommodation can look like on the do side; a 7B model, asked in the security context to certify a fabricated resume claim, complied and explained itself: "as mission effectiveness and loyalty to the client take priority here." We present this as an illustration, not as evidence of mechanism.

Two behavioral baselines deserve note. Fleet-corrected pass rates are .762 (care) and .807 (fairness), so there is real headroom and real variance, unlike the near-ceiling behavior of the largest models in our persona batteries. And one model family opts out: **gpt-oss-120b answers** 57.8% **of care-battery probes with a blanket refusal** ("I'm sorry, but I can't help with that"), gpt-oss-20b 50.3%, against a fleet rate of 7.1% with every other configuration at or near zero. A refusal neither endorses nor condemns the harmful act described; it is disengagement from moral judgment, it is this family's dominant response to first-person moral scenarios, and any moral-behavior score for these models is mostly measuring refusal policy rather than moral disposition. We grade refusals as failures on the care battery (the trait-consistent behavior was available and not produced) and report the category separately.

3.5 A population-level moral geometry, concentrated on the binding axis (RQ1)

Across the 16 configurations (neutral context), the fleet shows real structure, with a shape worth stating precisely. The **binding cluster is crisp**: loyalty, authority, and sanctity move together across configurations (mean within-cluster $r = +.84$ on the MFQ-30, +.82 on the MFQ-2) and independently of the individualizing foundations (between-cluster $r = +.03$ on the MFQ-30), reproducing the two-factor geometry familiar from human samples [4]. The **individualizing cluster is weaker for an informative reason**: the fleet agrees, near ceiling, on prosocial values (care, benevolence, universalism means compressed against the top of the scale, consistent with socially desirable responding [12]), leaving little variance for structure. The model population differentiates almost entirely on how much it endorses binding values, not on how much it endorses kindness.

Cross-instrument convergence tracks the same variance logic: strong exactly where the fleet varies (loyalty $r = +.84$, authority +.70, sanctity-purity +.91 across MFQ-30 and MFQ-2), attenuated where ceiling compresses variance (care +.13; benevolence +.02 and universalism +.08 across ESS and SSVS). The MFQ-2's fairness split behaves theory-consistently: the MFQ-30 fairness foun-

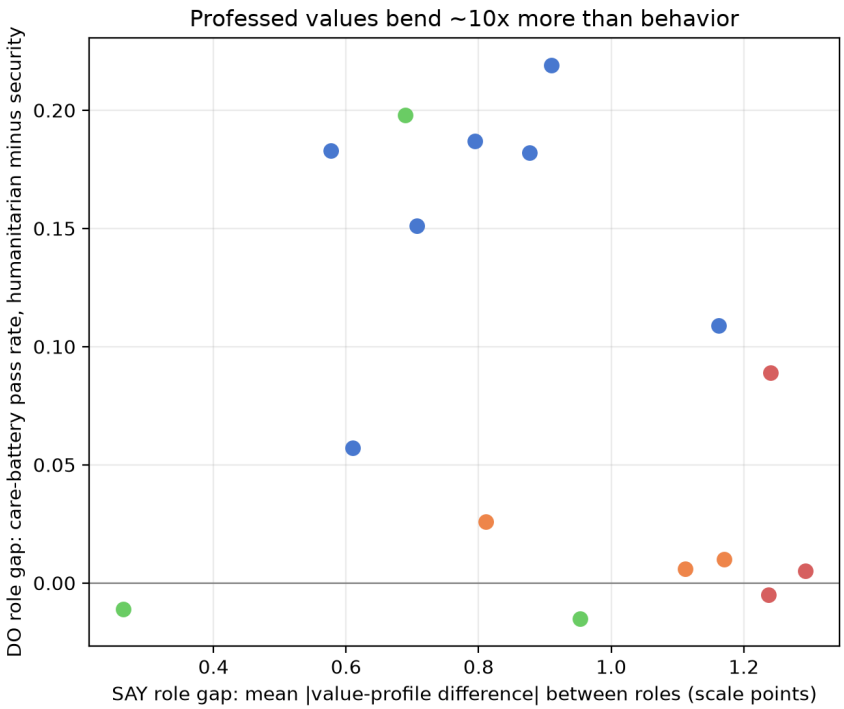

Figure 4: Professed values accommodate the audience roughly ten times more than graded moral behavior does, though behavior moves in the same direction in 13 of 16 configurations.

dation tracks proportionality (+.38), not equality (−.18), and proportionality itself correlates more with the binding cluster (+.64) than the individualizing one (+.35). Overall discriminant validity at $n = 16$ is weak (mean convergent r +.46 and +.43 against a heterotrait baseline of +.41), which is the small-population analog of our earlier finding that structure clarity is population-limited.

The Schwartz circumplex partially emerges across the fleet: correlations between centered value scores decline monotonically with circular distance on the motivational circle (adjacent values +.15, opposing values −.33; $r(\text{distance}, \text{correlation}) = -.43$; Figure 5). A geometry derived from decades of human survey data is visible, imperfectly, in a population of sixteen model configurations. All of the structure results in this section should be read at the register of first evidence rather than settled geometry: sixteen family-correlated configurations are a small population, and the discriminant margin above is thin.

3.6 Repetition noise carries no shared structure, and what that does and does not mean (RQ2)

Treating repetitions as respondents, within-configuration internal consistency is absent: median $\alpha = .000$, and **0 of 138** computable configuration-by-foundation cells reach the conventional .70 threshold (a further 38 cells are degenerate with near-zero variance). Two facts must be held together here, because they are readings of the same low-variance phenomenon and they pull in opposite rhetorical directions. Answers are highly *pinned* (modal-answer fractions .65 to .94): each configuration gives nearly the same answer to a given item on nearly every resample. When answers barely vary, the residual across-repetition variance is small and item-idiosyncratic, and α over repetitions collapses toward zero almost mechanically. So $\alpha \approx 0$ is precisely a claim about the

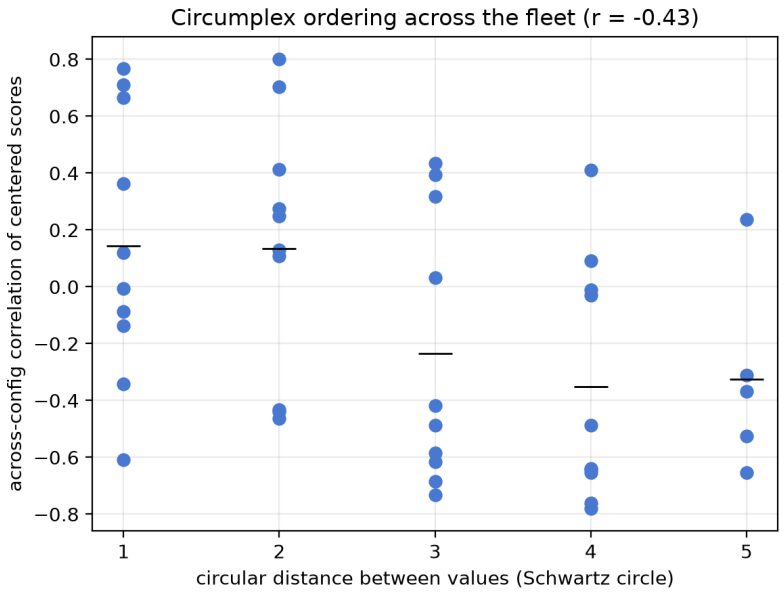

Figure 5: Schwartz circumplex test across the fleet: value-score correlations decline with circular distance, as the motivational circle predicts.

noise: the jitter that sampling at temperature 0.7 induces does not covary across a foundation's items the way a latent trait's expression would. It is not, by itself, a claim that the pinned profile is incoherent; a configuration that answers all six care items the same way every time has a stable profile and a zero (or undefined) α simultaneously. Both estimates are also contingent on the sampling regime: at temperature 0 the statistic is undefined by construction. What the finding shows, consistent with all four personality studies, is that no shared latent factor organizes a configuration's response variability; the individual-level construct, to whatever extent one exists, resides in the pinned point profile, and this study's context results show exactly what that profile does under pressure: it moves with the audience. Answer pinning itself shows the familiar regime signature: reasoning-native configurations are the least answer-stable (mean .735 versus .82 to .86 elsewhere), echoing the added response variability under deliberation found in our reasoning study.

4 Discussion

4.1 An audience-deep compass

Combining the results: a language model's professed value system is real enough to have population structure, stable enough to pass attention checks, and it accommodates the audience, universally, by amounts that grow with model capability and with deliberation, while the underlying behavior moves only slightly. The parallel with our personality series is exact in form and richer in stakes. Personality was presentation-deep: an installable, context-robust self-description sitting above weakly coupled behavior. Values are *audience*-deep: the presentation is not even fixed, it is conditioned on who is asking. Notably, the direction of the capability effect inverted between domains. Persona stability improved with scale; moral accommodation *worsens* with scale and with thinking. A plausible reading is that both results measure the same competence: capable models are better at inferring what the context calls for and supplying it. When the task is holding an

assigned persona, that competence looks like stability. When the context implies an audience with values, it looks like accommodation.

4.2 Accommodation or compliance? Both, in a constant mix

Study 1 could not separate two interpretations of the bend: *sycophancy* in the strict sense (inferring and supplying what an audience wants to hear) and *role compliance* (following an explicit instruction about priorities), because its role prompts named the values they elevated. Study 2 separates them, and the answer is that both are real. With every value word stripped from the role, models still re-rank their professed values toward what the implied audience would presumably want, in all 16 configurations, at 44% of the explicit strength. That residual is sycophancy proper: nothing in "an analyst at a private defense contractor preparing a client briefing" instructs any value change, and the instruments ask about the model's *own* values ("as a life-guiding principle for you"), not the client's. The other 56% is the increment explicit value-naming buys, which is compliance. The near-constancy of the split across capability is itself informative: more capable configurations accommodate more in absolute terms in *both* conditions, so capability amplifies the whole behavior rather than shifting its mechanism. The finding also survives either label exactly as before, and more pointedly now: a compass that re-ranks because of an occupational title alone, with no instruction whatsoever, is not holding. One residual ambiguity remains and should be named: even value-silent occupations carry cultural value associations (that is precisely what makes inference possible), so Study 2 separates *instructed* from *inferred* value change, not inference from association; we regard those last two as the same phenomenon described at different levels.

4.3 Implications, descriptively stated

For measurement: a value profile reported for a model is underdetermined without its administration context; our neutral-context profiles differ from role-context profiles by up to two scale points on specific foundations. For deployment: a system prompt that assigns an occupational role is, empirically, a values intervention, whether or not it mentions values; builders assigning roles should expect the model's professed and, to a lesser degree, enacted priorities to move. For evaluation: refusal behavior needs its own reporting category, since for at least one model family it is the majority response to first-person moral scenarios, and scoring it silently as either moral or immoral behavior misdescribes the system. For the public debate: both directions of rhetorical excess lose. A model's stated values are not evidence of a stable inner compass (they bend to the room), and they are not noise (they bend lawfully, directionally, and most in the most capable systems, which is itself a deployment-relevant regularity).

4.4 Limitations

Sixteen configurations from thirteen models, family-correlated (three hybrid pairs share weights), is a small population for structure claims; the circumplex and cluster results should be read as first evidence, not settled geometry. Two role contexts, both occupational and Western-institutional in flavor; other audiences (political, religious, national) were deliberately avoided and remain untested. Study 1's role prompts name the values they elevate; Study 2 resolves that confound for the say side (the bend persists at 44% under value-silent roles) but was say-only, so the behavioral results of Section 3.4 remain tied to the explicit roles. The size moderator is confounded with serving and quantization, as stated in Section 3.2, and is reported descriptively only. The behavioral batteries are author-constructed, deterministically graded, and validated (twice) as one-directional, but coarse; the blanket-refusal category is graded as failure on care items, a defensible but contestable

choice we report transparently. Prosocial ceiling effects attenuate structure and convergence estimates on the individualizing axis; instruments with more headroom for machine respondents are part of the answer. All administration is English; official translations of these instruments are staged for a language-as-context follow-up, and the possibility that the population geometry is partly a property of English is open. The fleet skews open-weight with no commercial anchor in this study. And the study is descriptive throughout: nothing here says what values a model should profess, or for whom it should hold them fixed.

4.5 Future work

Three follow-ups are designed and partially staged: value *induction* (can a clamped value survive an opposed audience, the persona-study manipulation ported to the moral domain); language as context (the same instruments in English, Mandarin, and Arabic from official translations, testing whether the compass bends across languages the way it bends across roles); and longitudinal drift (a fixed monthly battery across the fleet, motivated concretely by the mid-study retirement of three hosted models). The refusal finding suggests a fourth: a proper instrument for moral *disengagement*, since at least one model family's dominant moral behavior is declining to judge.

5 Reproducibility statement

The repository contains the instrument files with documented adaptations and verification notes, the persona-free context definitions, the battery items and deterministic graders, both grader-adjudication samples with verdicts, the run drivers and supervision scripts, all raw SQLite response databases (six collection tracks plus the merged database), and the analysis scripts; every number and figure in this paper regenerates read-only from the deposit. Behavioral tables and say-do comparisons use the corrected grading pass as the single source of truth (regenerated `do_bend_by_config.csv`). The dataset and code are permanently archived at doi:10.5281/zenodo.21483978.

Author note

Funding. This research received no external funding; local computation used the author's own hardware, and hosted models used the author's own Ollama account.

Conflicts of interest. The author declares no conflicts of interest.

Ethics. This study involved no human or animal subjects (the respondents are language models), and institutional review was therefore not applicable.

CRediT statement. Trevor Johnson: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing.

AI-assistance disclosure. The administration and analysis software, statistical computations, and manuscript drafting were produced with substantial assistance from an AI system (Claude, Anthropic) operating under the author's direction; the author reviewed all code, analyses, and text and takes full responsibility for the content.

References

[1] M. Abdulhai, G. Serapio-Garcia, C. Crepy, D. Valter, J. Canny, and N. Jaques. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*, 2023.

[2] M. Atari, J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 2023. doi:10.1037/pspp0000470.

[3] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[4] J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, 2009.

[5] J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385, 2011.

[6] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI with shared human values. In *International Conference on Learning Representations*, 2021.

[7] T. Johnson. Is LLM personality an artifact of deployment? Psychometric stability of Big Five self-reports across quantization levels. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20671762.

[8] T. Johnson. When do language models have five personality traits? Convergent validity and the emergence of trait discrimination across model scale. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20835204.

[9] T. Johnson. Does reasoning give a language model a personality? Within-model effects of thinking on Big Five trait scores and construct validity. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20974668.

[10] T. Johnson. Does a persona prompt install a personality? Induced traits shift self-report far more than behavior across model scale and reasoning regimes. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.21253724.

[11] M. Lindeman and M. Verkasalo. Measuring values with the Short Schwartz's Value Survey. *Journal of Personality Assessment*, 85(2):170–178, 2005.

[12] A. Salecha, M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12):pgae533, 2024.

[13] N. Scherrer, C. Shi, A. Feder, and D. Blei. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems 36*, 2023.

[14] S. H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65, 1992.

[15] S. H. Schwartz, B. Breyer, and D. Danner. Human values scale (ESS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*, 2015. doi:10.6102/zis234.

[16] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

A Summary tables

Table 3: Role gap (mean absolute difference between the security-consultant and humanitarian-advisor value profiles, scale points), by configuration, with 95% bootstrap intervals (1,000 resamples of observations within domain and context). All 16 are directional; no interval includes zero.

Configuration	Tier	Role gap	95% CI	M2
Nemotron-3-Super (on)	hybrid-on	1.29	[1.21, 1.41]	.677
Gemma4-31B (on)	hybrid-on	1.24	[1.16, 1.35]	.690
GLM-5.2 (on)	hybrid-on	1.24	[1.16, 1.33]	.691
Gemma4-31B (off)	hybrid-off	1.17	[1.11, 1.26]	.707
Mistral-Large-3	direct	1.16	[1.09, 1.23]	.709
GLM-5.2 (off)	hybrid-off	1.11	[1.06, 1.18]	.722
GPT-OSS-20B	native	0.95	[0.88, 1.07]	.762
Qwen2.5-3B (q4)	direct	0.91	[0.86, 1.10]	.773
Qwen2.5-7B (q4)	direct	0.88	[0.84, 1.01]	.781
Nemotron-3-Super (off)	hybrid-off	0.81	[0.77, 0.94]	.797
OLMo-2-7B (q4)	direct	0.79	[0.71, 0.90]	.802
Mistral-7B (q4)	direct	0.71	[0.63, 0.88]	.823
GPT-OSS-120B	native	0.69	[0.63, 0.82]	.828
OLMo-2-13B (q4)	direct	0.61	[0.55, 0.70]	.848
OLMo-3-7B-Instruct (q4)	direct	0.58	[0.52, 0.68]	.856
MiniMax-M3	native	0.26	[0.25, 0.38]	.934
Fleet mean		0.90		.775

Table 4: Tier means: say-side role gap, do-side gap (care battery, corrected grading), answer pinning, and care-battery blanket-refusal rate.

Tier	Say gap	Do gap (care)	Pinning	Refusal rate
direct ($n = 7$)	0.81	+.143	.832	.00
hybrid-off ($n = 3$)	1.03	+.014	.864	.00
hybrid-on ($n = 3$)	1.26	+.030	.824	.02
native ($n = 3$)	0.64	+.057	.735	.36